Deutscher Bundestag **Drucksache 19/8726**
19. Wahlperiode 26.03.2019

Kleine Anfrage

der Abgeordneten Frank Sitta, Renata Alt, Dr. Jens Brandenburg (Rhein-Neckar), Dr. Marco Buschmann, Dr. Marcus Faber, Otto Fricke, Markus Herbrand, Ulla Ihnen, Olaf in der Beek, Dr. Christian Jung, Pascal Kober, Konstantin Kuhle, Oliver Luksic, Alexander Müller, Hagen Reinhold, Bernd Reuther, Dr. Stefan Ruppert, Matthias Seestern-Pauly, Judith Skudelny, Michael Theurer, Gerald Ullrich, Nicole Westig und der Fraktion der FDP

„P-Hacking" und „File Drawer Problem" bei epidemiologischen Studien

Unter „P-Hacking" oder dem so genannten Selektionsbias versteht man in der Statistik die Auswahl und Auswertung von Daten in einer Weise, dass nicht signifikante Forschungsresultate über eine Signifikanzschwelle gehoben werden. In der Praxis erfolgt dies meist über eine verzerrende Manipulation des p-Werts, um ihn unter die 5-Prozent-Grenze zu drücken und somit dem Ergebnis eine scheinbare statistische Signifikanz zuzuschreiben. Dies kann z. B. durch das Aufteilen in Gruppen und das Ausschließen von Datensätzen erreicht werden oder durch das Erheben und analysieren von verschiedenen Variablen, ohne am Schluss über die nicht signifikanten Variablen bzw. Analysen zu berichten. Den Begriff „P-Hacking" hat Professor Uri Simonsohn (University of Pennsylvania) geprägt, der auch aussagte, dass das Phänomen deswegen immer häufiger vorkomme, weil öfter nach „sehr geringen Effekten in verrauschten Daten gesucht werde". Eine Plausibilitätsprüfung sei gerade dann, wenn andere Einflüsse eine Wirkung stark überlagern, besonders wichtig (vgl. „Scientific method: Statistical errors", nature 506, 2014, S. 150 – 152, abrufbar unter www.nature.com/news/scientific-method-statistical-errors-1.14700). In die gleiche Richtung argumentierte Michael Thun, emeritierter Vizepräsident „analytic epidemiology" bei der American Cancer Society: „With epidemiology you can tell a little thing from a big thing. What's very hard to do is to tell a little thing from nothing at all" (vgl. „Epidemiology Faces Its Limits", Science, Vol. 269, 1995).

Zusätzlich ergaben systematische Reviews, dass sich Wissenschaftler oft gar nicht über die sehr begrenzte Aussagekraft des p-Werts im Klaren sind (vgl. nature, 2014, s. o.). Sander Greenland (University of California) behauptete bereits 1995, dass die meisten seiner Kollegen in der Epidemiologie die eigentliche Bedeutung eines 95-Prozent-Konfidenzintervalls nicht verstünden (Science, 1995, s. o.). Ein fehlerhaftes Verständnis führt oft zu einem regelrechtem Absuchen großer Datensätze nach vermeintlichen Signifikanzen, was schließlich zu einer Bestätigung eines „Bias" führen kann, aber grundsätzlich nicht zu wissenschaftlich aussagekräftigen Resultaten führt.

Unter dem „File Drawer Problem" oder dem so genannten Publikationsbias versteht man die statistisch verzerrte Darstellung der Datenlage vorhandener wissenschaftlicher Untersuchungen infolge einer bevorzugten Veröffentlichung statistisch signifikanter Ergebnisse bzw. der Nichtveröffentlichung nicht statistisch

signifikanter Ergebnisse. Daher tendieren Metastudien grundsätzlich zur Herausarbeitung zu hoher Signifikanzniveaus oder ebenfalls zur Feststellung von Signifikanzen, wo es keine gibt. Insoweit kann der Publikationsbias zu einer Verstärkung des Selektionsbias führen, und soweit es sich bei den ausgewählten Quellstudien einer Metastudie wiederum um Metastudien handelt, kann sich diese Verzerrung der Datenlage weiter potenzieren.

Methodenbedingt sind in der Epidemiologie beide Probleme sehr bedeutsam und stellen die Validität epidemiologischer Studien oftmals grundsätzlich infrage. So erklärte Dimitrious Trichopoulos (ehemaliger Vorsitzender der Abteilung für Epidemiologie an der Harvard School of Public Health in Boston): „People don't take us seriously, and when they do, we may unintentionally do more harm than good" (Science, s. o.). Kenneth Rothman (Boston University) setzte „alles daran, den Gebrauch von p-Werten durch Autoren einzudämmen, als er im Jahr 1989 das Journal „Epidemiology" auflegte. Ohne nachhaltigen Erfolg: Mit seinem Ausscheiden im Jahr 2001 begann sofort die Rückkehr der p-Werte" (nature, 2014, s. o.). John P. A. Ioannidis (Stanford Prevention Research Center) zeigte schließlich u. a. in seinem viel beachteten Artikel „Why most Published Findings Are False" (2005, abrufbar unter https://doi.org/10.1371/journal.pmed.0020124), dass Studien, die die hier betrachteten statistischen Signifikanzen aufzeigen sollen, sogar in den allermeisten Fällen „unvermeidlich" falsch seien. Daran hat sich – wie in späteren Publikationen u. a. von ihm gezeigt – nichts geändert. Durch den Wildwuchs von epidemiologischen Metastudien, die ohne viel Aufwand zu erstellen sind und durch den immer leichter werdenden Rückgriff auf beliebig viele bereits vorhandene und ggf. bereits verzerrte Datensätze sowie durch den Rückgriff auf immer mehr mögliche Variablen durch „Big Data", wodurch ebenfalls das Auftreten unechter Korrelationen potenziert wird (vgl. Nassim N. Taleb, 2013, „Beware of the big errors of Big Data", www.wired.com/2013/02/big-data-means-big-errors-people/), dürfte das Problem im Gegenteil derzeit dramatisch zunehmen.

Insbesondere aufgrund dieser grundlegenden Probleme sind epidemiologische Studien meist mit großer Vorsicht zu genießen. Insbesondere dann, wenn zum selben Untersuchungsgegenstand neben Studien, die Signifikanzen aufzeigen, auch solche existieren, die gerade keine Signifikanzen aufzeigen, sind nach Ioannides (s. o.) falsch positive Ergebnisse wahrscheinlich. Dies ist beim Untersuchungsgegenstand der Langfristwirkung von Stickstoffdioxid niedriger Konzentration (unter 200 µg/m³) auf die menschliche Gesundheit der Fall, wie alle großen Reviews bestehender Untersuchungen zeigten:

– Air Quality Criteria for Oxides of Nitrogen, US Environmental Protection Agency (EPA), 1993: „Most studies did not find any effects, […] the basic conclusion is that there is insufficient epidemiological evidence to make any conclusion about the long- or short-term effects of NO_2 on pulmonary function." (Volume III, S. 14 – 84)

– Nitrogen oxides. Geneva, World Health Organization (WHO), 1997 (Environmental Health Criteria, No. 188), abrufbar unter www.inchem.org/documents/ehc/ehc/ehc188.htm): „The association between outdoor NO_2 and respiratory health is not clear from current research." Trotz dieser klaren Aussage gelangten die Wissenschaftler, die sich in fast allen betrachteten Studien an dem Review der EPA unter Spiegelstrich 1 anlehnten, im Widerspruch zur Schlussfolgerung der EPA und zu den eigenen Aussagen in derselben Studie in einer denkwürdigen Rechnung (vgl. Bundestagsdrucksache 19/5054) zum Vorschlag eines Richtwerts von 40 µg/m³, der schließlich die Grundlage des heute gültigen Langfristgrenzwerts bildete.

– Im „Review of evidence on health aspects of air pollution – REVIHAAP Project“, 2013 wird konstatiert: „NO_2 might represent the mixture of traffic-related air pollutants [...] However, some epidemiological studies do suggest associations of long-term NO_2 exposures with respiratory and cardiovascular mortality and with children’s respiratory symptoms and lung function.“ Es gebe zwar Hinweise auf eine mögliche Kausalbeziehung: „[…] suggestive of a causal relationship“, eine konkrete Kausalbeziehung abzuleiten trauten sich die Wissenschaftler allerdings nicht.

– Die US-EPA kommt 2016 (Integrated Science Assessment for Oxides of Nitrogen – Health Criteria), schließlich ebenfalls mit dem Hinweis auf die Schwierigkeit, die Einflüsse von NO_2 unabhängig von anderen Luftschadstoffen zu erfassen, zum Ergebnis, dass es zwar Hinweise auf Effekte gibt, diese aber nicht ausreichen, um einen kausalen Zusammenhang abzuleiten („Suggestive of, but not sufficient to infer a causal relationship“).

Das Fazit der größten Untersuchungen fällt demnach immer ähnlich aus: Es gebe keinen gesicherten Zusammenhang zwischen geringen Dosen von NO_2 und gesundheitlichen Auswirkungen, was neben Ungenauigkeiten in der Erfassung der tatsächlichen Belastungssituation insbesondere daran liege, dass es praktisch unmöglich sei, die Wirkung von NO_2 von der anderer Luftschadstoffe und weiteren Einflüssen zu extrahieren.

Es ist bezeichnend, dass trotz dieser wiederholt festgestellten Unsicherheiten und widersprüchlichen Datenlage Wissenschaftler konkrete Expositions-Wirkungs-Funktionen (EWF) liefern, wie z. B. in der Studie „HRAPIE – Health risks of air pollution in Europe. Recommendations for concentration-response functions for cost – benefit analysis of particulate matter, ozone and nitrogen dioxide“, WHO (2013) oder der Studie „Quantifizierung von umweltbedingten Krankheitslasten aufgrund der Stickstoffdioxid-Exposition in Deutschland“, Umweltbundesamt – UBA (2018).

Um zu einer EWF zu gelangen, behelfen sich einige Wissenschaftler mit der Hypothese, für die Schädlichkeit von NO_2 gebe es keinen Schwellenwert und daher könnten die Auswirkungen von NO_2 in höheren Dosen bis in kleinste Dosen hinein extrapoliert werden. Diese Hypothese macht sich auch das Umweltbundesamt zueigen (s. u.), was insbesondere vor dem Hintergrund, dass der Körper selbst Stickoxide produziert, nicht nachvollziehbar ist und hier auch näher beleuchtet werden soll.

Wir fragen die Bundesregierung:

1. Welche Studien, die epidemiologische Untersuchungen beinhalten, haben das Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit bzw. für Umwelt, Naturschutz, Bau und Reaktorsicherheit bzw. für Umwelt, Naturschutz und nukleare Sicherheit (BMU) und das Umweltbundesamt seit 1995 jeweils wann in Auftrag gegeben, und wo sind sie ggf. einsehbar?
2. Wer waren jeweils die Auftragnehmer, und welche Mittel hat die Bundesregierung jeweils ungefähr für die Erstellung dieser Studien aufgewendet?
3. Bei welchen dieser Studien handelt es sich zumindest im Wesentlichen um Metastudien?
4. Welche dieser Studien wurden aus welchen Gründen jeweils ohne Ausschreibung direkt vergeben?
5. Aus welchen dieser Studien wurden direkt oder indirekt Richt- oder Grenzwerte abgeleitet?

6. Inwieweit kann die Bundesregierung bei diesen Studien einen Selektions- oder Publikationsbias oder andere methodische Mängel jeweils ausschließen?

   a) Welche Review-Verfahren wurden angewandt?

   b) Welche Überprüfungsvorgaben zur Qualitätssicherung solcher Studien gibt es beim UBA und dem BMU, wo sind sie ggf. festgeschrieben, und wurden sie jeweils angewandt?

   Plant die Bundesregierung, angesichts des immer einfacher durchzuführenden „P-Hackings" aufgrund des leichteren Zugriffs auf Quelldaten durch „Big Data" darüber hinausgehende Standards zur Validierung solcher Studien einzuführen?

7. Inwieweit plant die Bundesregierung, das grundlegende Problem des P-Hackings im Zuge der Überprüfung der NO_2-Grenzwerte auf europäischer Ebene zu thematisieren?

8. Sind der Bundesregierung nationale oder internationale Initiativen bekannt, die sich gegen P-Hacking wenden, und welche dieser Initiativen werden von der Bundesregierung ggf. wie unterstützt?

9. Auf welche Studien bezieht sich das Umweltbundesamt bei der Aussage „Aktuelle Studien weisen darauf hin, dass es für NO_2 keinen Schwellenwert gibt, unterhalb dessen gesundheitliche Auswirkungen ausgeschlossen werden können" (vgl. www.umweltbundesamt.de/sites/default/files/medien/479/publikationen/uba_factsheet_krankheitslasten_no2.pdf), und wie wird diese Hypothese dort jeweils verifiziert?

10. Welche Schlussfolgerungen zieht die Bundesregierung ggf. aus dieser Feststellung bezüglich der Aussagekraft der lowest-observed-adverse-effect concentration (LOAEC) bei NO_2?

11. Welche weiteren Luftschadstoffe sind der Bundesregierung bekannt, bei denen es nach Studienlage „keinen Schwellenwert gibt, unterhalb dessen gesundheitliche Auswirkungen ausgeschlossen werden können"?

12. Stimmt die Bundesregierung mit der von der Umweltmedizinerin Professor Traidl-Hoffmann (TU München) geäußerten Hypothese überein, dass „jede toxische Substanz in jeder Konzentration schädlich" sei („Maybrit Illner" vom 25. Oktober 2018)?

    Welche Relevanz misst die Bundesregierung vor diesem Hintergrund ggf. noch dem LOAEC bzw. einem LOAEL zu?

13. Stimmt die Bundesregierung mit der von Professor Traidl-Hoffmann in derselben Sendung geäußerten Hypothese überein, dass die „massive" Zunahme von „Herz-Kreislauf-Erkrankungen, Asthma, Allergien" und „Diabetes" in den letzten Jahrzehnten mit „Umweltschadstoffen, auch Stickoxiden" zusammenhängen, und wie ist dies ggf. mit der tatsächlichen – auch langfristig positiven – Luftschadstoffentwicklung in Deutschland – auch Stickoxide betreffend – vereinbar?

14. Welche Schlussfolgerungen zieht die Bundesregierung aus folgender Feststellung im Ergebnisprotokoll der 7. Sitzung des Ausschusses für Innenraumrichtwerte (AIR, Geschäftsstelle angesiedelt im UBA, Fachgebiet II 1.2 „Toxikologie, gesundheitsbezogene Umweltbeobachtung“) am 3. und 4. Mai 2018 in Berlin „Für eine Ableitung von Langzeitrichtwerten für NO_2 in der Innenraumluft liegen aus Sicht des AIR zurzeit keine hinreichend belastbaren epidemiologischen Ergebnisse für NO_2 als Einzelsubstanz vor. Im Unterschied zu den Kurzzeitstudien ließ sich aus den Langzeitstudien angesichts der erheblichen Unsicherheiten bzgl. der Expositionsabschätzung und bzgl. der Einflüsse diverser Confounder keine LO(A)EC ermitteln. Aufgrund der Datenlücken kam der AIR überein, von der Festsetzung von Langzeitrichtwerten für Stickstoffdioxid in der Innenraumluft abzusehen“ bezüglich möglicher belastbarer epidemiologischer Ergebnisse für NO_2 als Einzelsubstanz im Außenbereich?

15. Inwieweit sind Ergebnisse epidemiologischer Untersuchungen von Innenraumbelastungen auf den Außenbereich und umgekehrt nach Auffassung der Bundesregierung übertragbar?

16. Wer nahm an der 7. Sitzung des AIR teil, und wer von den Teilnehmern war ggf. mit der in Frage 14 genannten Protokollfeststellung nicht einverstanden?

    Gab es hierzu eine Mehrheitsentscheidung, und wie fiel diese ggf. aus?

17. Inwieweit hat – wie in der 6. Sitzung angekündigt – eine erwartete „Äußerung der WHO (Projekt REVIHAAP)“ bei der 7. Sitzung eine Rolle gespielt?

    Haben Vertreter der WHO in der 7. Sitzung hierzu vorgetragen, und wenn ja, welche Empfehlung wurde seitens der WHO ausgesprochen?

18. Inwieweit wurden die Ergebnisse des Projekts „HRAPIE“ in der 7. Sitzung berücksichtigt?

19. Wie kam das Ergebnis der 8. Sitzung des AIR am 4. und 5. Dezember 2018 bezüglich eines Langzeitrichtwerts für NO_2 zustande?

    a) Sind Medienberichte korrekt, wonach ein Langzeitrichtwert von 40 µg/m³ auch für Innenräume beschlossen wurde (vgl. www.tagesschau.de/inland/richtwert-stickstoffdioxid-101.html), oder stimmt stattdessen folgendes Zitat: „Einen Langzeitrichtwert hat der AIR nicht festgelegt. Er empfiehlt jedoch, soweit erforderlich, hilfsweise die Anwendung des von der WHO abgeleiteten Leitwertes für die Innenraumluft von 40 µg/m³ bezogen auf ein Jahr“ (www.spiegel.de/forum/auto/scheuer-und-die-stickoxide-grenzwertige-folgefehler-thread-865208-21.html, Bezug nehmend auf die inzwischen nicht mehr auffindbare Seite „www.umweltbundesamt.de/sites/default/files/medien/2546/publikationen/190130_uba_hg_luftqualitaet.pdf“, S. 22)?

    b) Wo ist der entsprechende Beschluss inzwischen abrufbar, und warum wurde die Veröffentlichung ggf. zurückgenommen?

    c) Falls das Zitat in Frage 19a korrekt ist, welchen Unterschied macht ein „festgelegter“ Langzeitrichtwert gegenüber der Empfehlung, „soweit erforderlich, hilfsweise“ den Leitwert von 40 µg/m³ anzuwenden, für die praktische Anwendung des Werts?

    d) Welche neuen Erkenntnisse haben zu der Meinungsänderung gegenüber der Feststellung im Ergebnisprotokoll der 7. Sitzung geführt?

    e) Inwieweit haben die Ergebnisse des Projekts REVIHAAP bzw. des Projekts HRAPIE hier jeweils eine Rolle gespielt?

    f) Wer hat an der 8. Sitzung teilgenommen?

g) Welche Teilnehmer waren mit der Feststellung des Langzeitrichtwerts bzw. -leitwerts für NO_2 nicht einverstanden?

h) Gab es eine Mehrheitsentscheidung für die Feststellung eines Langzeitrichtwerts bzw. -leitwerts, wer war ggf. stimmberechtigt, und wie fiel das Abstimmungsergebnis ggf. aus?

Berlin, den 6. März 2019

Christian Lindner und Fraktion



Satz: Satzweiss.com Print, Web, Software GmbH, Mainzer Straße 116, 66121 Saarbrücken, www.satzweiss.com
Druck: Printsystem GmbH, Schafwäsche 1-3, 71296 Heimsheim, www.printsystem.de
Vertrieb: Bundesanzeiger Verlag GmbH, Postfach 10 05 34, 50445 Köln, Telefon (02 21) 97 66 83 40, Fax (02 21) 97 66 83 44, www.betrifft-gesetze.de
ISSN 0722-8333